MULTIPLE REGRESSION ANALYSIS OF AVERAGE BROWSING TIME

STAT 431 Final Project

Abstract

A multiple regression analysis was performed to predict the average browsing time of a user using predictors from the Mozilla Labs Test Pilot Study conducted in 2010. Predictors considered were 1) number of windows open, 2) number of tabs, 3) number of bookmarks, 4) % of private sessions, 5) number of downloads, 6) number of extensions, and 7) number of browser activations. The model considering 1), 2), and 7) was found to be statistically significant. The most notable result of the model is a parabolic relationship between browsing time and number of windows, suggesting that browsing time declined significantly after two windows are opened.

Bill Guo

STAT 431 Fall 2015

1. Introduction

Today, through surveillance and data collection, everyone from advertising agencies and web companies to the government knows how we browse the Internet. In fact, for the most part, they know more about how we access the Web than we do. This is especially concerning when we consider that most of us spend most of our time on the Internet every day.

In this study, browsing data is analyzed in order to build a multiple regression model for the timespan of an individual's browser session. Predictors considered are number of windows, number of tabs, % of private sessions, number of downloads, number of extensions, and number of browser activations. Through this exercise we hope to gain some insight as to how people behave online, and learn more about the activity that has so rapidly become an integral part of today's society.

2. Dataset Details

2.1 Data Collection

The dataset analyzed was provided by Mozilla Labs, a part of the Mozilla, and consists of browser usage patterns (Mozilla Firefox, specifically) for 27,000 users recorded during a week in November 2010[1]. Mozilla Labs is a place of innovation and experimentation with Web technologies, so naturally this data was interesting to them in designing products for the Web in the future.

Notably, the data was collected for users that voluntarily installed the Test Pilot extension. It is likely that such users are not representative of the browsing habits of the general population. Going forward, we will need to keep this in mind.

2.2 Data Description

The dataset has three parts. First is a table of 27,000 users, stored anonymously as numeric IDs. For each user, Firefox version, operating system, and number of extensions was stored. In this analysis, only the user IDs and number of extensions were used. Notably, while it is assumed that each user ID corresponds to a different person, this is not necessarily the case. If someone had multiple computers running Firefox or multiple Firefox profiles open, they could contribute several submissions and thus represent multiple IDs. As there is nothing we can do about this situation, the purposes of this study every ID is assumed to match a unique individual.

The second table is an enormous dump of browser events recorded for each user over the period of a week. Browser events included occurrences such as browser launch/browser shutdown, download completion, memory usage, etc. Every event is timestamped and matched to a user. The events analyzed in this study were:

1. BROWSER_START / BROWSER_SHUTDOWN: These recorded launch and shutdown of the browser. The timestamp difference was used in order to obtain a value for browsing time, which is the response variable we are seeking to model.
2. BROWSER_ACTIVATE: This event corresponds to when a browser stops idling. Idling is defined by the lack of active use for 10 minutes. The number of browser activations is used as a predictor variable. We predict a positive correlation with browsing time, or that the more activations, the longer the average browsing session.
3. BOOKMARK_STATUS: This event was recorded on startup, and took down the number of bookmarks saved. This variable was of some interest, as by many accounts bookmarks do not seem to be a widely used feature anymore. We tentatively predict a negative correlation, assuming that people save bookmarks when they want, but don't

have the time, to fully investigate a site. On the other hand, perhaps people who browse for longer may stumble onto more interesting sites that could call for a bookmark.

4. DOWNLOAD: This event was recorded when a download was completed. We predict a positive correlation, as people who browse more would naturally have more time to find and download files.
5. PRIVATE_ON / PRIVATE_OFF: These events recorded when a user entered/exited private browsing. The effects of this variable do not appear obvious before analysis.
6. NUM_TABS: This event recorded both the number of windows open and the number of tabs open in the browser session. It is recorded at launch and every 15 minutes thereafter. We predict a positive correlation since more time spent browsing allowed for more time to open tabs and windows.

Several other events were also recorded, but these were taken to be of primary interest to the study.

The final table contained answers to a survey on browsing habits, with questions such as where said user accessed the Internet most frequently or time spent on the Web daily. However, the survey was optional, so plenty of data was missing. As a result, this table was ignored.

Data Usage and Processing

Due to the extreme size of the events portion of the dataset, a subset of 402 users and their recorded events was used instead. This was also provided by Mozilla Labs. Even after cutting out ~26500 users there were 997,176 events recorded. For some context, the size of the unfiltered events table in .csv format is 3.6 GB; the smaller subset is 50 MB.

In order to obtain values for the response variable, browsing session length, we took the timespan difference between browser startup and browser shutdown. Here there was evidently some issues with Mozilla's data collection, as there were more startups than shutdowns recorded. To account for this, we only used recorded BROWSER_SHUTDOWN events that were immediately followed by a recorded BROWSER_STARTUP, since no events should be recorded between those two. All other instances of the two events, excepting the initial recorded startup and the final recorded shutdown, were discarded as anomalously recorded.

The more likely explanation is that the user never closed the browser during the week long recording period. However, there is no way to account for this and obtain a true browsing session time, so these were simply tossed out.

After this processing, only 369 of 402 user IDs remained with browsing sessions that started and ended during the data collection period. Despite the filtering, 369 users is still a large enough sample to work with and obtain meaningful conclusions.

Events for these 369 users were pulled from the table and processed into a separate .csv file for analysis:

Response:

1. Average browsing session time (btime): This is the response variable. The lengths of browsing sessions during the 1 week recording period for each user were averaged to obtain these values. It is in minutes.

Predictors:

1. Average number of windows (n_windows): This was obtained by recording the average of the windows opened data from NUM_TABS events.
2. Average number of tabs (n_tabs): Also obtained from NUM_TABS.

3. Average number of bookmarks extant (n_bookmarks): Average of BOOKMARK_STATUS.
4. % of private sessions (priv): This was recorded from PRIVATE_ON / PRIVATE_OFF as the number of private sessions opened, and then divided by the number of sessions opened by the user in total.
5. Downloads per session (dls): This was recorded from DOWNLOAD to obtain the number of downloads in total, then divided by the number of sessions in total.
6. Number of extensions (ext): This was extracted from the users table.
7. Activations per session (acts): This was recorded directly from BROWSER_ACTIVATE events, and divided by number of sessions in total.

Hereafter, for brevity, each variable will be referred to by the name in parentheses.

3. Data Analysis

3.1 Initial Diagnostics

The first step taken was to check the distributions of each variable. This was done using histograms, in order to check for extreme outliers and gross errors.

One value was in fact found to be highly anomalous. The average browsing time recorded was 84717.32 minutes, which is roughly 54 days – impossible given that the collection period was only a week. The only explanation is faulty data collection, so it was thrown out. Now there are 368 instead of 369 rows in our set.

Scatter plots between btime and the predictors were made to further check for outliers and also for potential transformations of the predictors. Speaking generally, there were many outliers with extremely high values for browsing time. None other than the erroneous one were discarded.

It was found that a log transformation of btime linearized its relationships with n_windows, n_tabs, and n_bookmarks. Therefore, a log transformation of the response was taken. Although this is generally not advisable, it seemed appropriate given that the transformation linearized half of the predictors.

A log transformation was considered for acts, as it was highly positively skewed, and it seemed likely that the log would linearize its relationship to log(btime). However, there were 32 users who had 0 activations. To bypass this issue, the transformation log(acts+.1) was taken instead.

Multicollinearity was preliminarily checked for using a scatterplot matrix (Fig 1). From knowledge of the variables, there might be a linear relationship between n_windows and n_tabs. However, the matrix did not seem to provide strong evidence. After fitting a model, VIFs will be examined to come to better conclusions.

3.2 Data Partitioning

Generally, at this step it is customary to divide the data into training and test sets, in order to avoid overfitting. However, the size of this dataset is only 368 rows. It seemed ill advised to perform the partitioning because of this small sample size. According to statisticians, data splitting is a costly method in terms of reducing the accuracy of the model's predictions, and generally makes the model unreliable[2]. This was also seen during the analysis when trying to fit models using different partitions. With this in mind, models were fitted on the entire dataset.

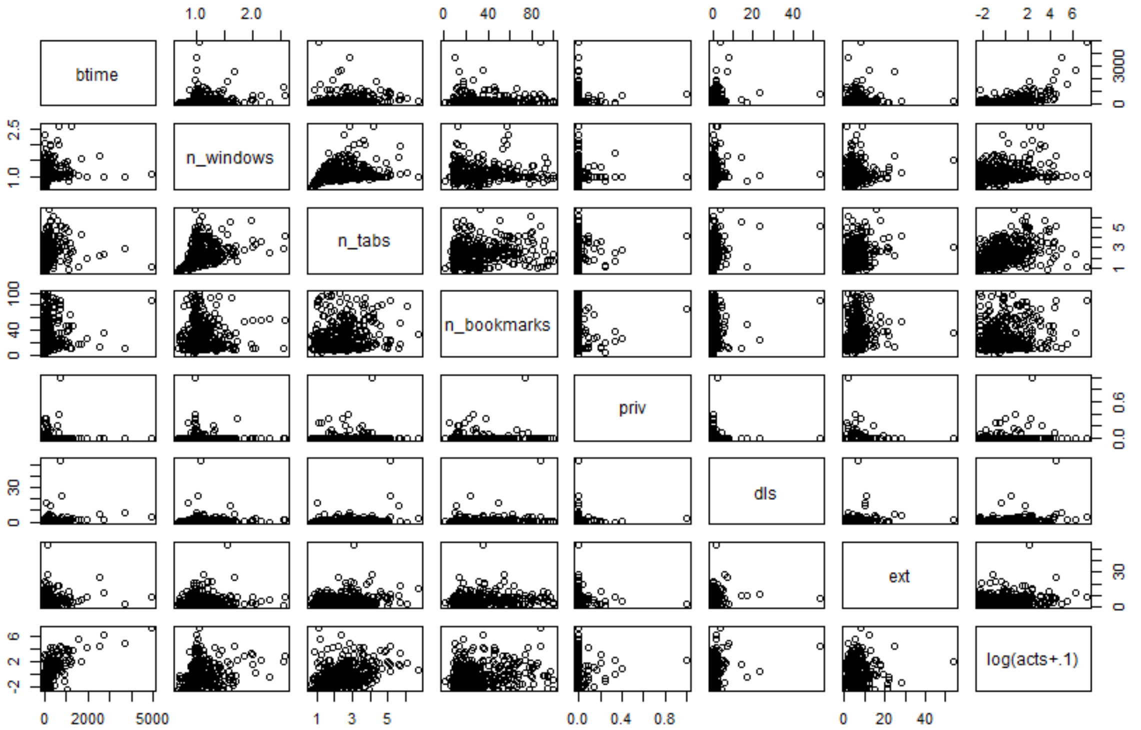


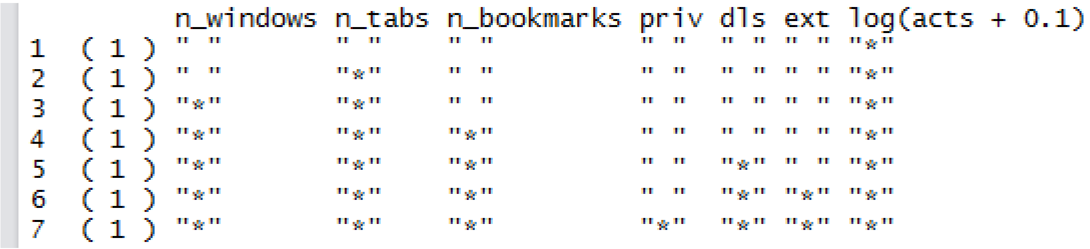
Figure 1 - Scatterplot matrix of all variables. No strong collinearity observed.

3.3 Fitting Candidate Models

Using stepwise regression in both directions, a candidate model was proposed.

Candidate model: $\log(btime) = \beta_1 n\_windows + \beta_2 n\_tabs + \beta_3 \log(acts + .1)$

Next, best subsets regression was used to fit more candidate models. This was possible due to the relatively low number of predictors under consideration. Eight candidate models were proposed (Fig 2).



```
         n_windows n_tabs n_bookmarks priv dls ext log(acts + 0.1)
1  ( 1 ) " "       " "    " "         " "  " " " " "*"
2  ( 1 ) " "       "*"    " "         " "  " " " " "*"
3  ( 1 ) "*"       "*"    " "         " "  " " " " "*"
4  ( 1 ) "*"       "*"    "*"         " "  " " " " "*"
5  ( 1 ) "*"       "*"    "*"         " "  "*" " " "*"
6  ( 1 ) "*"       "*"    "*"         " "  "*" "*" "*"
7  ( 1 ) "*"       "*"    "*"         "*"  "*" "*" "*"
```

Figure 2 - Best subsets regression output.

Notably, every model included log(acts+.1). This is expected, since activations should be highly related to length of browser session. We also note that the proposed stepwise regression model is included in the best subsets results, as the 3rd in the list.

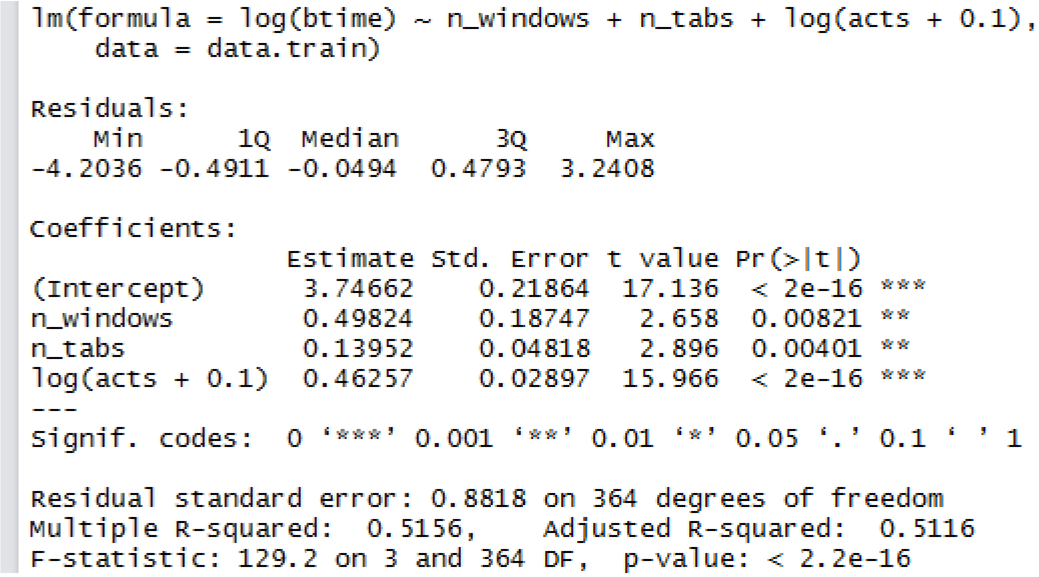
Using the variable selection criterion, 3 models were chosen for further evaluation. Ranked by AIC, the top 3 models were #3, #2, and #4, in order. These rankings were the same using BIC and using Mallow's C_p.

Model 1: $\log(btime) = \beta_0 + \beta_1 n\_windows + \beta_2 n\_tabs + \beta_3 \log(acts + .1)$

Model 2: $\log(btime) = \beta_0 + \beta_1 n\_tabs + \beta_2 \log(acts + .1)$

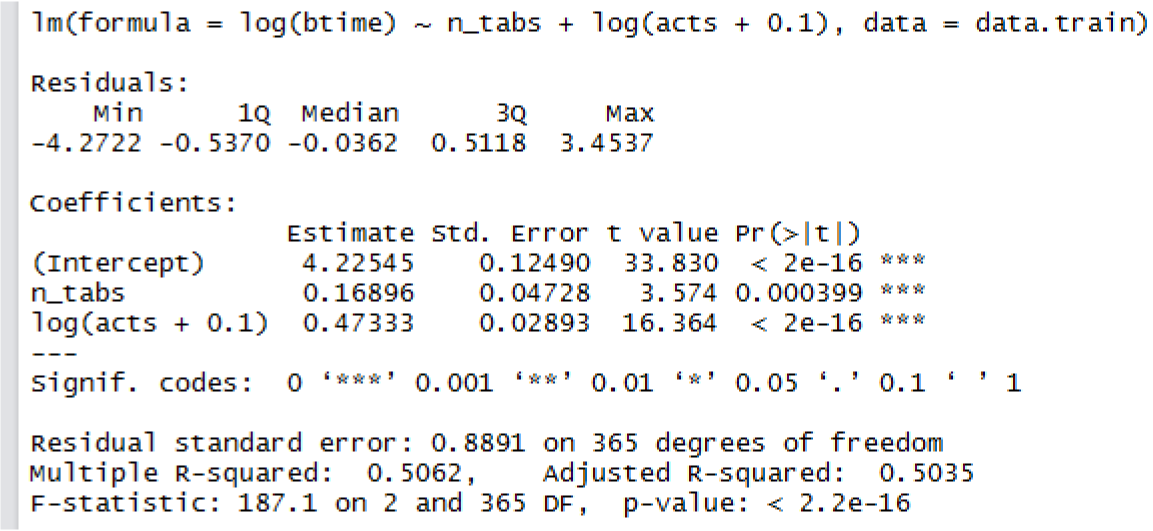
Model 3: $\log(btime) = \beta_0 + \beta_1 n\_windows + \beta_2 n\_tabs + +\beta_3 n\_bookmarks +$ $\beta_4 \log(acts + .1)$

R summaries for the calculated coefficients, standard errors, and P-values of each predictor for each model are depicted below (Fig 3, 4, 5).



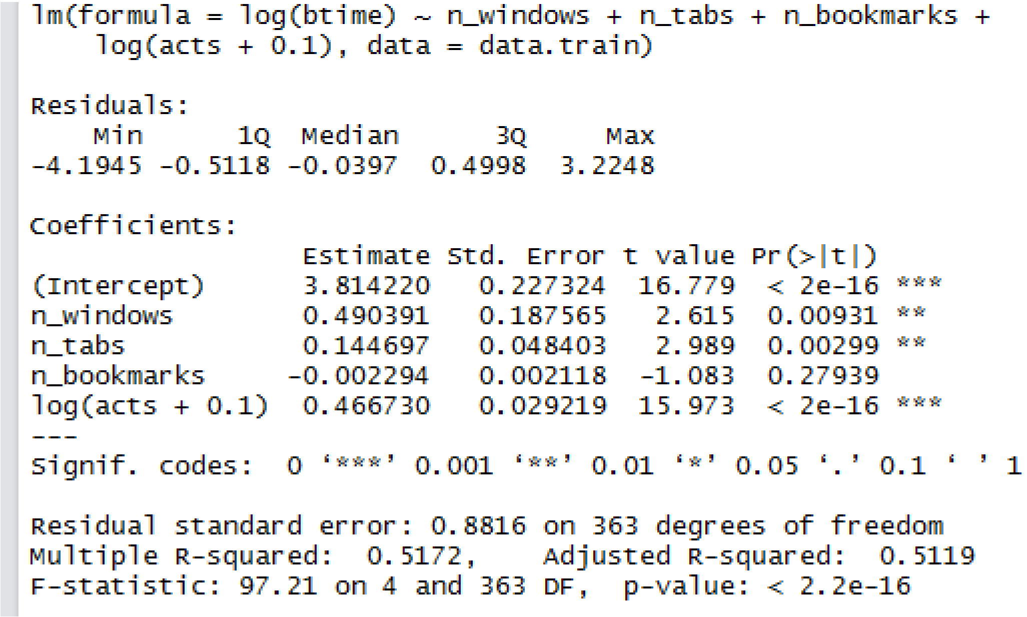
```
lm(formula = log(btime) ~ n_windows + n_tabs + log(acts + 0.1),
    data = data.train)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.2036 -0.4911 -0.0494  0.4793  3.2408 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      3.74662    0.21864  17.136  < 2e-16 ***
n_windows        0.49824    0.18747   2.658  0.00821 ** 
n_tabs           0.13952    0.04818   2.896  0.00401 ** 
log(acts + 0.1)  0.46257    0.02897  15.966  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8818 on 364 degrees of freedom
Multiple R-squared:  0.5156,	Adjusted R-squared:  0.5116 
F-statistic: 129.2 on 3 and 364 DF,  p-value: < 2.2e-16
```

Figure 3 - Model 1



```
lm(formula = log(btime) ~ n_tabs + log(acts + 0.1), data = data.train)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.2722 -0.5370 -0.0362  0.5118  3.4537 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      4.22545    0.12490  33.830  < 2e-16 ***
n_tabs           0.16896    0.04728   3.574 0.000399 ***
log(acts + 0.1)  0.47333    0.02893  16.364  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8891 on 365 degrees of freedom
Multiple R-squared:  0.5062,	Adjusted R-squared:  0.5035 
F-statistic: 187.1 on 2 and 365 DF,  p-value: < 2.2e-16
```

Figure 4 - Model 2



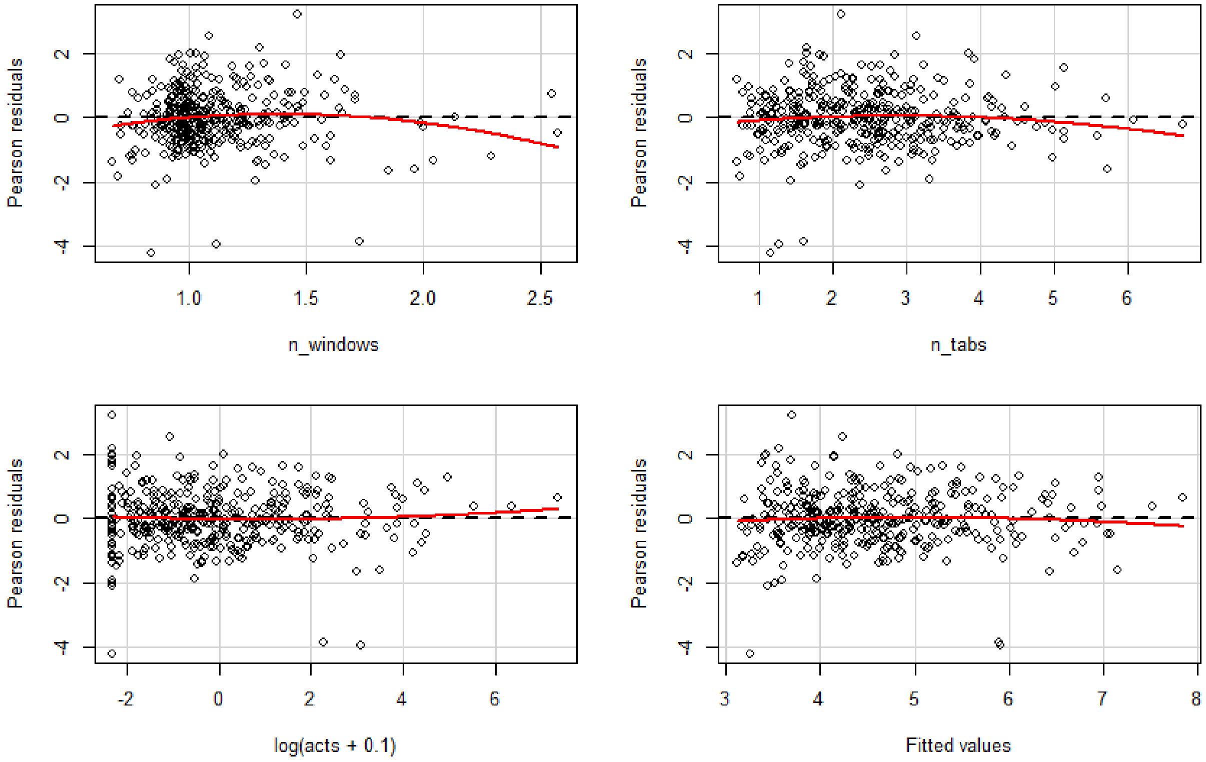
```
lm(formula = log(btime) ~ n_windows + n_tabs + n_bookmarks + 
    log(acts + 0.1), data = data.train)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.1945 -0.5118 -0.0397  0.4998  3.2248 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      3.814220   0.227324  16.779  < 2e-16 ***
n_windows        0.490391   0.187565   2.615  0.00931 ** 
n_tabs           0.144697   0.048403   2.989  0.00299 ** 
n_bookmarks     -0.002294   0.002118  -1.083  0.27939    
log(acts + 0.1)  0.466730   0.029219  15.973  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8816 on 363 degrees of freedom
Multiple R-squared:  0.5172,	Adjusted R-squared:  0.5119 
F-statistic: 97.21 on 4 and 363 DF,  p-value: < 2.2e-16
```

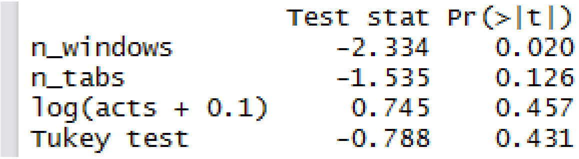
Figure 5 - Model 3

Looking at the output for Model 3, the P-value for n_bookmarks is too high; we cannot reject the null hypothesis that n_bookmarks is not part of the model with 95% confidence. Therefore we discard Model 3.

Furthermore, given that Model 2 differs from 1 only in the removal of n_windows, and since logically n_windows should be somewhat related to browsing time, we omit any discussion of Model 2. We proceed with diagnostics on Model 1.

Plotting the residuals against the fitted values, there is no significant evidence of heteroscedasticity. However, there is significant evidence on nonlinearity that needs to be fixed, made clear by the residual plots against each variable (Fig 6).





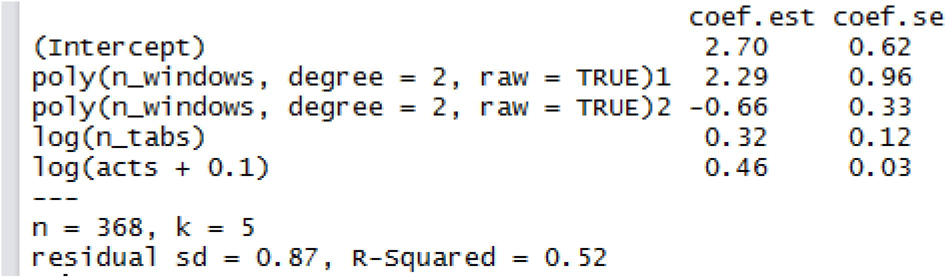
```
                  Test stat Pr(>|t|)
n_windows            -2.334    0.020
n_tabs               -1.535    0.126
log(acts + 0.1)       0.745    0.457
Tukey test           -0.788    0.431
```

Figure 6 - Partial residual plots and curvature tests for Model 1.

The P-values for n_windows and n_tabs are too low to reject curvature. In order to fix this, we refit the model using polynomial terms.

By taking the 2nd-order polynomial of n_windows and the log of n_tabs, the nonlinearity is eliminated. The summary and diagnostic plots for the resulting Model 1.1 are depicted (Fig 7, 8, 9). The resulting model 1.1:

$$\text{Model 1.1: } \log(btime) = 2.70 + 2.29n\_windows - .66n\_windows^2 + .32\log(n\_tabs) + .46\log(acts + .1)$$



```
                                              coef.est coef.se
(Intercept)                                    2.70     0.62  
poly(n_windows, degree = 2, raw = TRUE)1       2.29     0.96  
poly(n_windows, degree = 2, raw = TRUE)2      -0.66     0.33  
log(n_tabs)                                    0.32     0.12  
log(acts + 0.1)                                0.46     0.03  
---
n = 368, k = 5
residual sd = 0.87, R-Squared = 0.52
```

Figure 7 - Model 1.1

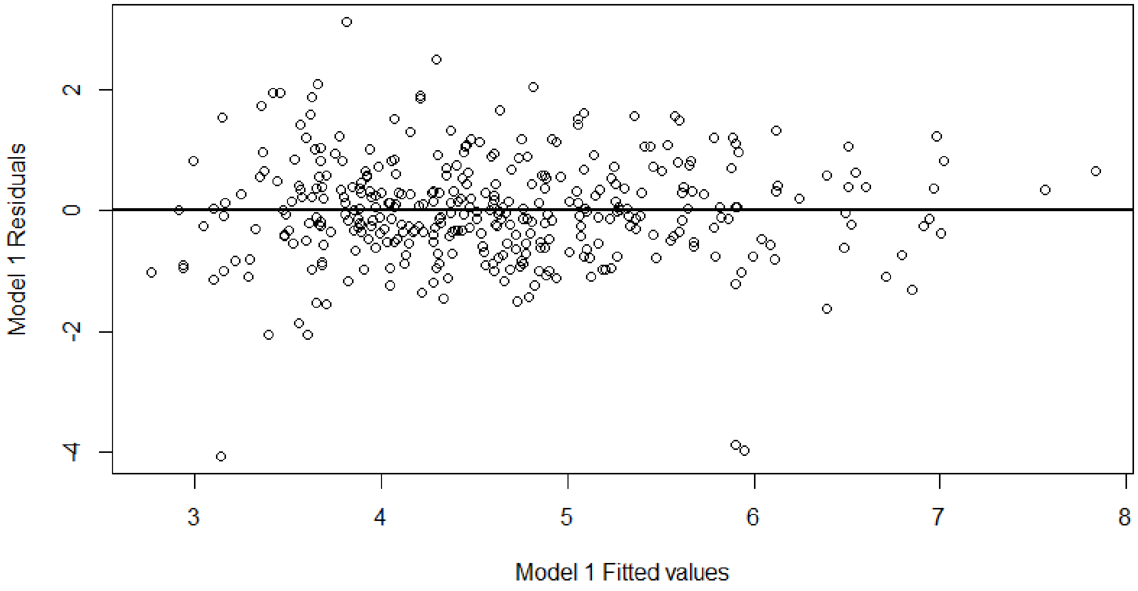
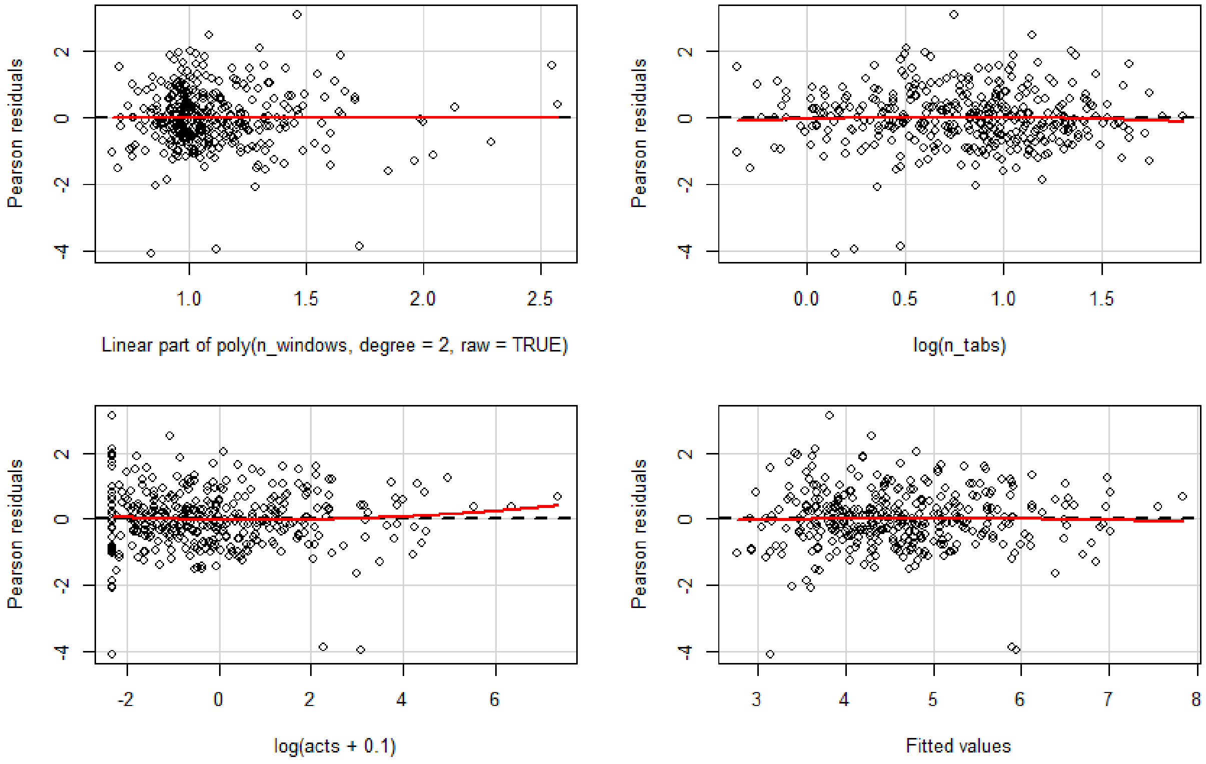
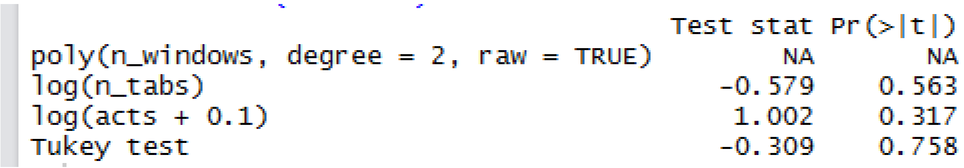


Figure 8 Residuals v. Fitted values for Model 1.1. No significant heteroscedasticity.





```
                                        Test stat Pr(>|t|)
poly(n_windows, degree = 2, raw = TRUE)        NA       NA
log(n_tabs)                                -0.579    0.563
log(acts + 0.1)                             1.002    0.317
Tukey test                                 -0.309    0.758
```

Figure 9 - Partial residual plots and curvature tests for Model 1.1.

The next diagnostic performed was on normality of the residuals, using a normal QQ plot. It is depicted in Figure 10.

A singular heavy tail is noticed, which could pose problems for the predictive accuracy of our fitted model. Namely, prediction intervals will be inaccurate. However, the negative effects of non-normality can generally be ignored using robust methods, so the heavy tail was left in the model. As suggested by Tamhane and Dunlop, non-normality is the least important of the assumptions[3].

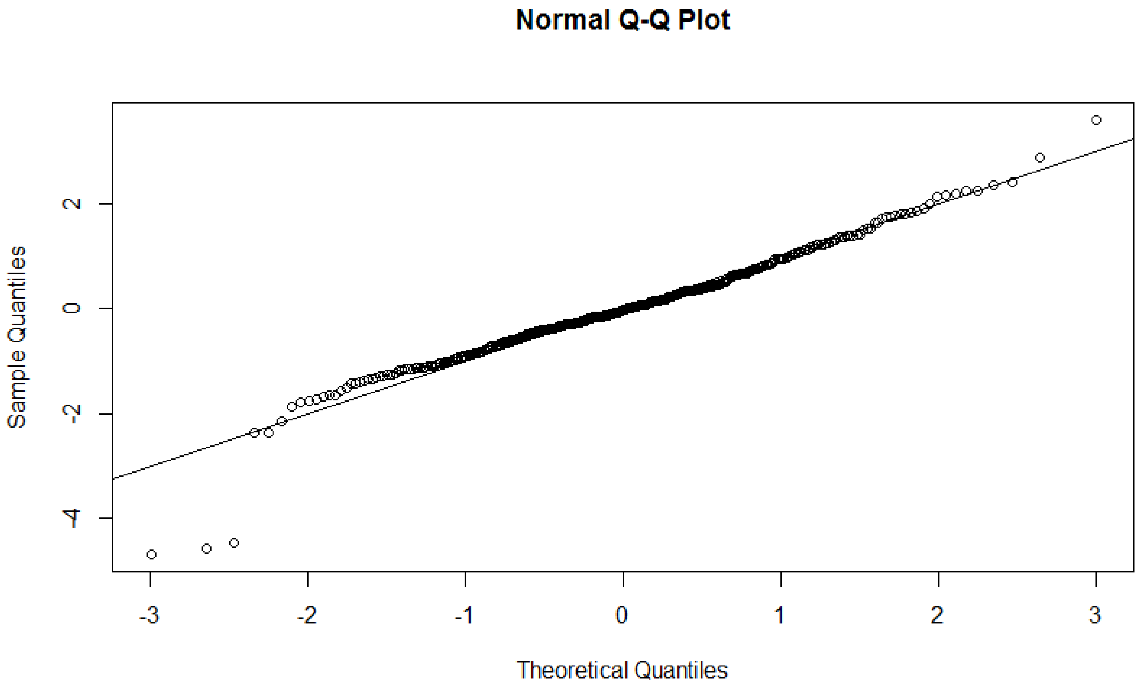


Figure 10 - QQ plot of residuals for Model 1.1. A heavy tail is observed.

We then check the hat values and Cook's distance to find influential observations (Fig 11, 12). Using both plots, several influential observations are noticed. While the Cook's distance never surpasses 1, that rule of thumb does not seem to hold, as clearly there are several observations with Cook's distances that are orders of magnitude larger than the rest. These observations represent users with browsing session times that are much longer than would be anticipated by the model; they will be discussed further in the results.

Finally, we check again for collinearity issues. As previously discussed, there could be a linear relationship between n_windows and n_tabs. The variance inflation factors are calculated for each predictor, to measure the effect of collinearity (Fig 13). The VIFs are low, suggesting little to no effect.

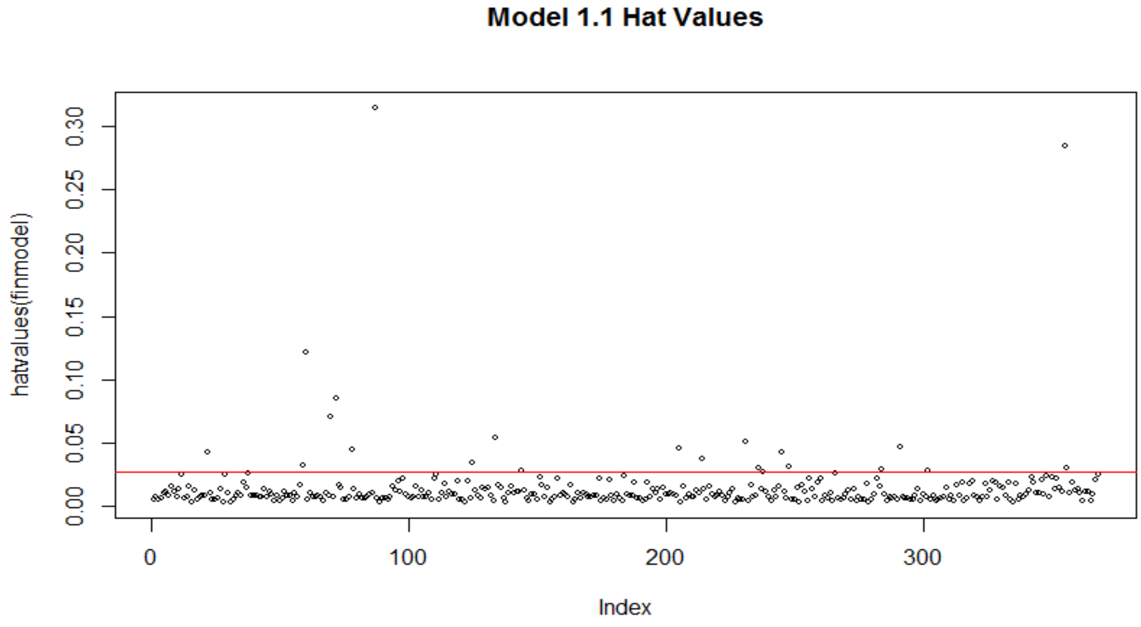


Figure 11 - Hat values for Model 1.1.

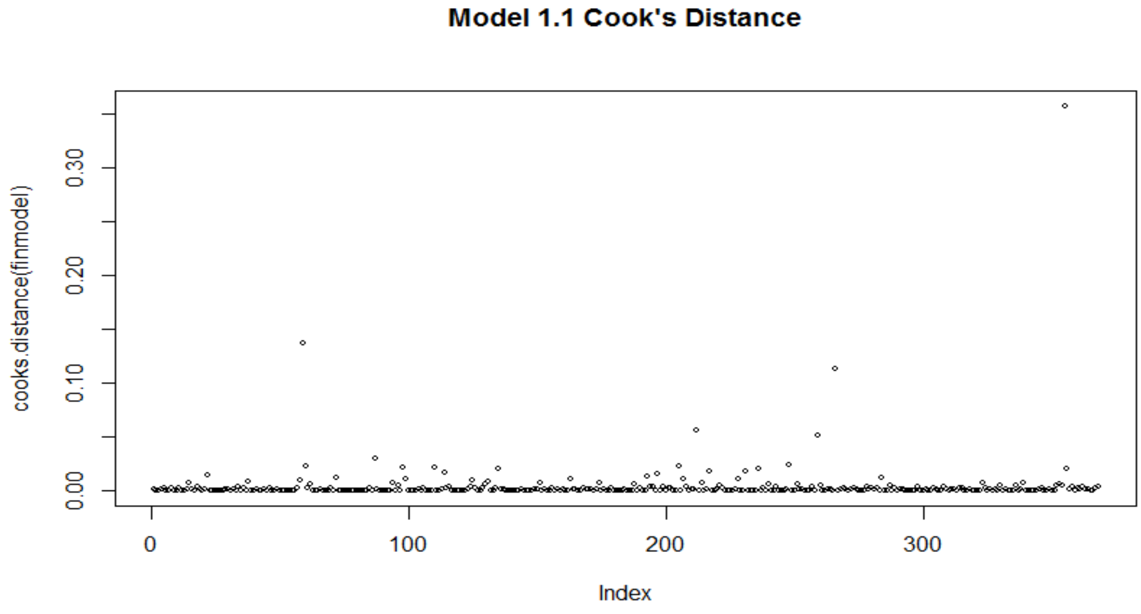
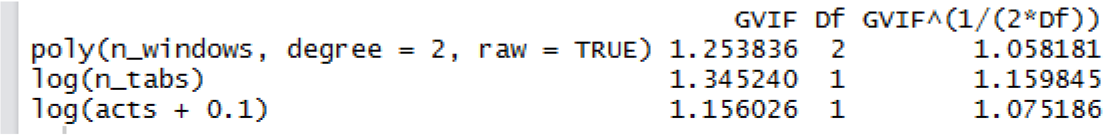


Figure 12 - Cook's Distance for Model 1.1.



```
                                            GVIF Df GVIF^(1/(2*Df))
poly(n_windows, degree = 2, raw = TRUE) 1.253836  2        1.058181
log(n_tabs)                             1.345240  1        1.159845
log(acts + 0.1)                         1.156026  1        1.075186
```

Figure 13 - VIFs for Model 1.1.

4. Results

The final model is:

$$\log(btime) = 2.70 + 2.29n\_windows - .66n\_windows^2 + .32\log(n\_tabs) + .46\log(acts + .1)$$

If we take e^x on both sides of the equation, we obtain the following:

$$btime = 14.8797\,(acts + .1)^{0.46} e^{(2.29n\_windows - 0.66n\_windows^2)} n\_tabs^{0.32}$$

We will analyze the form of the dependence of btime on each predictor.

Firstly, activations are related a $(acts + .1)^{.46}$. Of course, the result is positive, which is expected. More interestingly, notice that this function grows quickly at small values of acts, but then peters off, indicating that at high numbers of activations, extra activations do not contribute much to the browsing time. The interpretation is that people who idle many times do not differ from one another as much in browsing time, while people who idle sparingly have much more varied browsing times.

This is somewhat counterintuitive, because the only restriction on the time between activations is >10 minutes of idling. Activations could be 10 minutes, or 30, or 24 hours apart – acts does not care about the chronological distance. Logically, a user with more activations is more likely to have activations that are chronologically far apart, which would reduce the effectiveness of acts as a predictor of browsing time. Therefore we would expect the opposite result. Our tentative explanation is that this is simply due to the fact that our dataset contains more users with low numbers of activations than high.

There is a positive relationship between the number of tabs and the browsing time. This also makes sense, as longer browsing sessions allow for users to open more tabs. The form of the relationship, $n\_tabs^{.32}$, is similar to that of activations. In other words, browsing time grows quickly with n_tabs at low n_tabs values, but the growth rate declines rapidly at high n_tabs. This relationship makes sense. A browsing session with only one tab is much more likely to be short than a browsing session with even a slightly higher number of tabs. However, the difference between a browsing session with 20 tabs and 21 tabs is minimal.

The last predictor is n_windows, with the relationship $e^{2.29n\_windows-.66n\_windows^2}$. Graphing this relationship reveals a downwards-opening parabola with a maximum value around n_windows ~ 2. The function quickly decays to 0 around n_windows ~ 4. Approximate values of this factor are ~5 for n_windows = 1, ~7 for n_windows = 2, ~2 for n_windows = 3, and ~0 for n_windows = 4. This is an interesting result. One explanation is that attachments and documents often open in new windows, and users who need to open attachments and documents often only need to quickly print or read them before exiting. Another might have something to do with advertisements, which also often pop up in new windows. Perhaps it is easier to simply terminate the browser rather than close multiple intrusive ads manually. More thought and investigation is required to explain this peculiar finding.

5. Limitations

There are several limitations to this study. First and foremost is the sample size. While 368 rows was enough to build a model and produce results, a larger sample size would allow us to cross-validate our model via data splitting as well as improve the fit. This is especially true when we consider that the analyzed dataset was itself a subset of a larger dataset. While it proved too difficult to use the larger set in this exercise, future studies should certainly make use of the full data.

Another limitation is the users themselves. Our results may be significantly impacted by the fact that the reported data comes from users who voluntarily installed Mozilla's Test Pilot extension. Simply being aware of said extension suggests that these users are more technologically capable, which may influence how they browse in relation to the general populace. Unfortunately, it may be difficult to obtain a dataset on browsing habits for general users. Though such datasets likely exist in the databases of companies like Google and Microsoft, they may not be made public. Still, the conclusions of studies using those sets would be much more valid in general than this one.

A third limitation is the number of predictors considered in the model. Other predictors that could have been useful include operating system version, browser version, etc. While not used in this study, future projects would do well to consider the effects of other variables.

6. Conclusions

Ultimately, this study produced the following model for average browsing time, after eliminating 4 other candidate predictors:

$$\log(btime) = 2.70 + 2.29n\_windows - .66n\_windows^2 + .32\log(n\_tabs) + .46\log(acts + .1)$$

Fractional power dependencies were found between browsing time and both activations and number of tabs. This could be satisfactorily explained in the case of the number of tabs, but not in the case of the number of activations. Furthermore, the most notable conclusion from this model is the quadratic relationship of the number of windows and the browsing time, which suggested that browsing time peaks at 2 windows and declines sharply thereafter.

Future studies should examine these relationships in greater detail. In the case of activations, the hypothesis that the counterintuitive relationship found is simply a result of the small sample size can easily be verified or disproved. In the case of the number of windows, future studies should seek to confirm or disprove the result found, and provide a more compelling explanation.

7. References

1 Mozilla Labs. (2010). *Test Pilot: A Week in the Life of a Browser - Version 2: Aggregated Data Samples*. Retrieved 19 November 2015, from https://testpilot.mozillalabs.com/testcases/a-week-life-2/aggregated-data.html

2 Harrell, F. (2001). *Regression Modeling Strategies* (p.112). New York: Springer.

3 Tamhane, A., & Dunlop, D. (2000). *Statistics and Data Analysis: from Elementary to Intermediate* (p.369). Upper Saddle River, NJ: Prentice Hall.